Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers

Year: 2022 | Citations: 339 | Authors: C. Gao, F. Howard, N. Markov, E. Dyer, S. Ramesh

---

Abstract

Background Large language models such as ChatGPT can produce increasingly realistic text, with unknown information on the accuracy and integrity of using these models in scientific writing. Methods We gathered ten research abstracts from five high impact factor medical journals (n=50) and asked ChatGPT to generate research abstracts based on their titles and journals. We evaluated the abstracts using an artificial intelligence (AI) output detector, plagiarism detector, and had blinded human reviewers try to distinguish whether abstracts were original or generated. Results All ChatGPT-generated abstracts were written clearly but only 8% correctly followed the specific journal's formatting requirements. Most generated abstracts were detected using the AI output detector, with scores (higher meaning more likely to be generated) of median [interquartile range] of 99.98% [12.73, 99.98] compared with very low probability of AI-generated output in the original abstracts of 0.02% [0.02, 0.09]. The AUROC of the AI output detector was 0.94. Generated abstracts scored very high on originality using the plagiarism detector (100% [100, 100] originality). Generated abstracts had a similar patient cohort size as original abstracts, though the exact numbers were fabricated. When given a mixture of original and general abstracts, blinded human reviewers correctly identified 68% of generated abstracts as being generated by ChatGPT, but incorrectly identified 14% of original abstracts as being generated. Reviewers indicated that it was surprisingly difficult to differentiate between the two, but that the generated abstracts were vaguer and had a formulaic feel to the writing. Conclusion ChatGPT writes believable scientific abstracts, though with completely generated data. These are original without any plagiarism detected but are often identifiable using an AI output detector and skeptical human reviewers. Abstract evaluation for journals and medical conferences must adapt policy and practice to maintain rigorous scientific standards; we suggest inclusion of AI output detectors in the editorial process and clear disclosure if these technologies are used. The boundaries of ethical and acceptable use of large language models to help scientific writing remain to be determined.